Ontology based Machine Learning using Data Mining Techniques

Name:	**Mr. Manish Kumar Khare**
Guide Name:	**Dr. Arpita Gopal**
Proposal Title:	**Ontology based machine learning using data mining techniques**
PhD Id:	**100452**
Faculty:	**Computer Studies, SIU**
Email:	**meetmanishkhare@gmail.com**

Abstract

Learning, like intelligence, covers such a broad range of processes that it is difficult to define precisely. With respect to machines, it can be broadly said that a machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such a manner that it's expected future performance improves. The objective of this proposal is to build an e-healthcare framework that is discovered by converging the semantic web and web mining techniques. This e-healthcare framework would empower the patient and physician both for taking better decision for managing health. Healthcare institutions need to target the patients with proper portfolios of services. There is a wealth of data available within the healthcare systems but they lack effective analysis tools to discover hidden relationships and trends in data. By applying web mining techniques on ontology based structure of web, useful patterns and knowledge can be discovered for making intelligent decisions for the care of individuals.

Keyword: Ontology, Semantic web, Web mining, E-Healthcare, Association, Classification

Introduction

The healthcare environment is generally 'information rich' but 'knowledge poor'. There is a wealth of data available within the healthcare systems but they lack effective analysis tools to discover hidden relationships and trends in data. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences. Thus healthcare institutions need to target the patients with proper portfolios of services.

Valuable knowledge can be discovered with the application of mining technique to equip patients with knowledge for managing health for disease specific better care and understanding

through E-Healthcare. E-Healthcare is conceptually connecting the patients and doctors, medical information services and providing a platform to facilitate better understanding and care. The distinction between e-Healthcare and traditional healthcare is that the later can be termed as a doctor centered approach where the doctor gives consultation after analyzing the available information, past history and constraints of the patient based on his knowledge and experience. Hence doctor is the prime source of advice to the patient regarding his health problem, future care and medical prescription whereas “eHealth is a consumer-centered model of health care where stakeholders collaborate utilizing information and communication technologies (ICTs) including Internet technologies to manage health, arrange, deliver, and account for care, and manage the health care system”. Thus e-Healthcare facilitates Patient education for better health anytime and anywhere with the flexibility of time and place.

Motivation & Significance of the Problem

With respect to healthcare industry, there is a wealth of hidden information available that is largely untapped. The maximum percentage of data on the web are so much **unstructured, heterogeneous, distributed and time variant** that they can only be understood by humans, however the amount of data is gigantic that it can only be processed efficiently by machines. Therefore the great success of the current form of WWW leads to new challenges.

- **Data Interpretation using Machines:** A huge amount of data available online is interpretable by humans only, machine support in terms of interpretation of data available online is very much limited.
- **Intelligent Data Processing:** Since the size of the data in web is enormous it can only be processed by machine efficiently. However efficiency does not guarantee the intelligence; the machine support in terms of processing the data intelligently for determining useful patterns and knowledge is limited.

- **Personalization of the Information:** It is very much likely that people differ in the contents and presentations they prefer while interacting with the Web, more precisely Web organizers must know what the patients do and want. Furthermore it can be categorized in keeping track of individual preferences and facilitating recommendations as well.

The motivation behind this proposal is to address these challenges in the E-Healthcare sector to turn patient specific data into knowledge and determine useful pattern that can enable health practitioners and patients to make intelligent decisions.

Objective

Basically this proposal aims at reaping the benefits of the two fast developing research areas i.e. semantic web and web mining techniques by discovering a framework that integrates both the research areas. The key idea is to build a framework for the intelligent decision making in healthcare sector by converging the concept & techniques of semantic web and web mining. Some of the important objectives are enumerated below.

- E-Healthcare for Patients
- Knowledge Management for patients profile and diseases
- Recommendation to the subject expert (doctor) and patient based on the clinical test information
- Active delivery of information (based on ontologies and data mining techniques) creates a dynamic healthcare environment.
- Facilitating most cost effective way for delivering information and treatment.
- Linking of knowledge items (medical reports) with ontology for the preparation of patient specific treatment.

Methodology

By converging the semantic web and web mining techniques e-Healthcare evolve as a consumer-centered model of health care where various stakeholders collaborate utilizing information and communication technologies (ICTs) including Internet technologies to manage and care of individual's health. Following are the important factors that that equip patients with the essential knowledge for managing health to aid them in better conducting their health.

Education(eD) through online information: The concept of e-Health facilitates a change in the attitude of the patient with masses of information available on the Internet. The patients expose themselves to new information and possibly new value systems from individuals who put this information online.

Empowerment (eM) through medical history: Access to the knowledge base of medicine and medical history records with clinical test information has empowered patient and healthcare provider relationship. Patient has access to the same knowledge base that is accessed by healthcare provider. This makes as a compulsion for the physician to regularly update the relevant information to offer better health services that gives a consumer a new sense of empowerment.

Encouragement (eN) for the refined relationship between patient & physician

Extending (eX) scope of health care beyond its conventional boundaries: e-Health suppots the patients to obtain better care, prevent medical mistakes and have saved their own lives at times. These aspects have basically extended the scope of health care beyond its conventional boundaries making it a tool in the hands of the patient which they can utilize as knowledge for managing health.

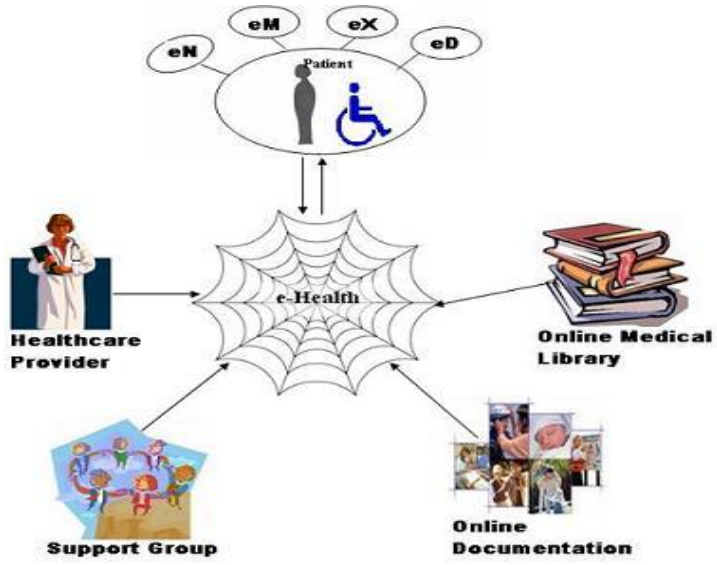

Figure 1: Knowledge Management in Healthcare

The knowledge management empower health care sector in predicting future health risks and educating patients online.

Proposed Framework

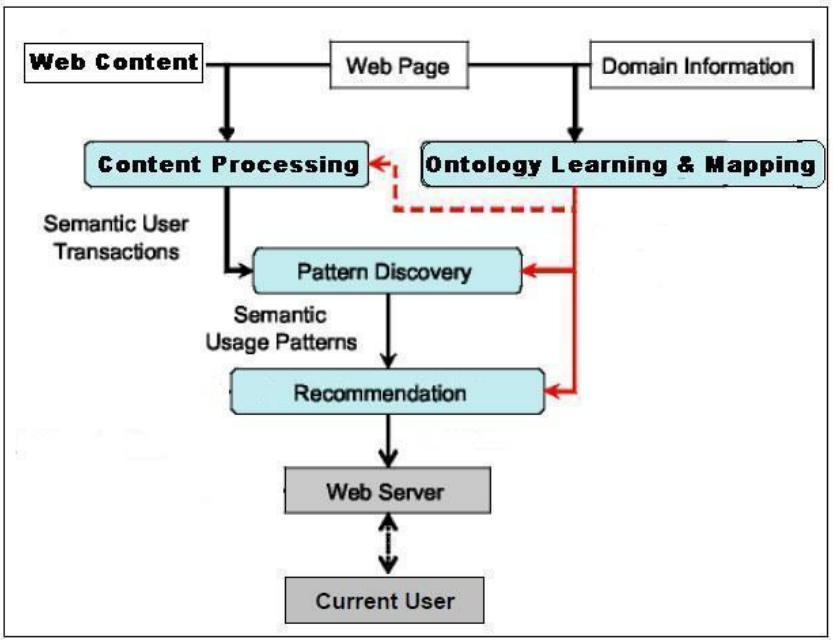

Figure 2: E-Healthcare Framework

Following are some example of health care cases that can avail vital information using this proposed framework.

- High number of count for White Blood Cell (WBC) : Outlier analysis
- Elderly woman suffering from diabetes leading to Miscarriage: Association technique
- Fever Categorization: Classification technique

Following tools can be used for building the suitable ontology structure for the proposed E-Healthcare system.

Karlsruhe Ontology Framework (KAON) : It is an open source ontology management and learning infrastructure targeted for business applications. It includes a comprehensive tool suite allowing easy ontology creation supported by machine learning algorithms.

FrameLogic: They provide a visual metaphor for representing the conceptual structure.

Description logic (DLs) is one most popular framework at the moment. DLs are subsets of first order logic which aim at being as expressive as possible while still being decidable.

Prot´eg´e-2000 is a platform-independent environment for creating and editing ontologies and knowledge bases. Like KAON, it has an extensible plug-in structure.

Sesame is architecture for efficient storage and expressive querying of large quantities of RDF(S) data.

Semantic Web

Introduction

The data available on Web is so much unstructured, heterogeneous and time variant that it is only interpretable by human mind and the machine support for this purpose is very limited. The

father of WWW, Mr. Tim Berners Lee suggests enriching the web by machine understandable information which supports the user in his tasks for interpretation. The Semantic Web addresses this challenge by trying to make the data machine-understandable as well. Current search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine understandable information can point the search engine to the relevant pages and can thus improve both precision and recall. The semantic Web is a step towards WI. It is based on languages that make more of the semantic content of the webpage available in machine-readable formats for agent based computing. One of the components of semantic Web techniques is to use ontology for marking up Web resources and assisting the generation and processing of semantic markup.

Ontology:

Ontology is an explicit formalization of a shared understanding of a conceptualization. It consists of following four terms.

Conceptualization: abstract model extracted from real world to represent some application

Explicit: It has explicit definition for concepts, relations, properties, functions & constraints

Formal: It can be understood & executed by machine

Share: Reflects commonly recognized domain knowledge

It provides a shared and common understanding of a domain that can be communicated between people and heterogeneous and widely spread application systems. It has a similar functionality as database schema with following differences.

- Ontology language is syntactically and semantically richer than common approaches for databases.
- It consists of semi-structured natural language texts and not tabular information.
- Ontology must be a shared because it is used for information sharing and exchange.
- Ontology provides a domain theory and not the structure of a data container.

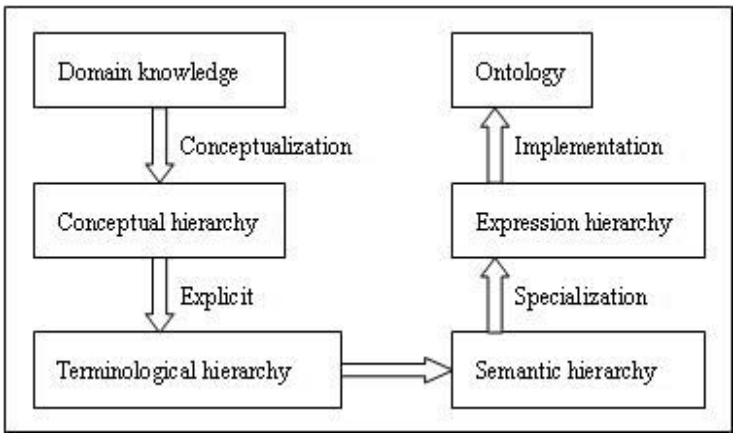

Figure 3: Ontology Creation Process

Following are the guideline for building the Ontology.

- Identification of the aim and the scope of the ontology
- Reuse existing vocabularies
- Enumerating most important term in the ontology
- Defining the classes and their hierarchies
- Defining the properties of the classes
- Defining the features of the properties
- Creating Instances

Layered Architecture of Semantic Web

Following figure shows the layers of the semantic web suggested by Mr. Tim Berner's Lee.

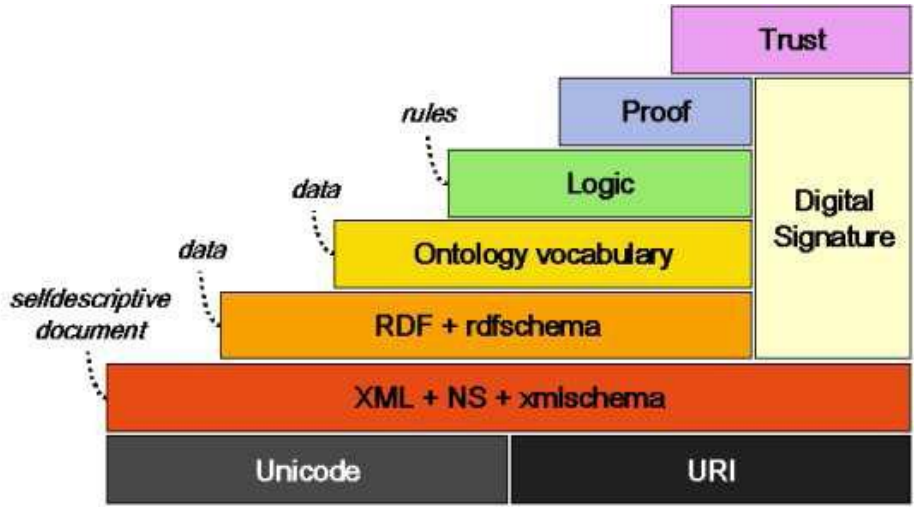

Figure 4: Layers of Semantic Web

Uniform Resource Identifiers (URI): provides a standard way to refer to entities

Unicode: It is a standard for exchanging symbols

Extensible Markup Language (XML): Provides syntax for structured document output

Namespaces: XML document can refer to the different namespaces to make explicit the context of different tags.

Resource Descriptor Framework: It is the first layer where information becomes machine understandable. According to the W3C recommendation, RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine understandable information on the Web." RDF documents consist of three types of entities: Resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any (real-world) objects which are not directly part of the WWW. In RDF, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource forms an RDF statement. A value is either a literal, a resource, or another statement. Statements can thus be considered as object-attribute value triples. In short RDF represents semantics of the web page contents.

Logic: This layer facilitates the retrieval of data as per the user request.

Ontology Language & Tools

Following are the names and details of Ontology language and tools.

Karlsruhe Ontology Framework (KAON) is an open source ontology management and learning infrastructure targeted for business applications. It includes a comprehensive tool suite allowing easy ontology creation supported by machine learning algorithms, ontology

management, and building ontology-based applications. The tool suite is also connected to databases to allow working with a large number of instances.

Frame Logic (or FLogic) is one candidate, since it provides a semantically founded knowledge representation based on the frame-and-slot metaphor. Another formalism that fits well with the structure of RDF is Conceptual Graphs. They also provide a visual metaphor for representing the conceptual structure.

Description logic (DLs) is one most popular framework at the moment. DLs are subsets of first order logic which aim at being as expressive as possible while still being decidable.

Prot´eg´e-2000 is a platform-independent environment for creating and editing ontologies and knowledge bases. Like KAON, it has an extensible plug-in structure.

Sesame is architecture for efficient storage and expressive querying of large quantities of RDF(S) data. It provides support for concurrency control, independent export of RDF(S) information, and a query engine for RQL, a query language for RDF

In section 4, the authors have also facilitated the practical aspect of extracting semantics from the Web. The authors have described the Semantic creation by Content & Structure with the description of Ontology learning, mapping and merging.

Web Mining

Web mining is the application of data mining techniques to extract knowledge from Web data, where **at least one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without** other types of Web data).

Web Mining is a triple **WM= (D, P, $).**

where D is the set of Web documents and activities; P is the set of patterns and $ is the mapping between D→P

Web mining classification is illustrated in the following diagram with the respective techniques in each category.

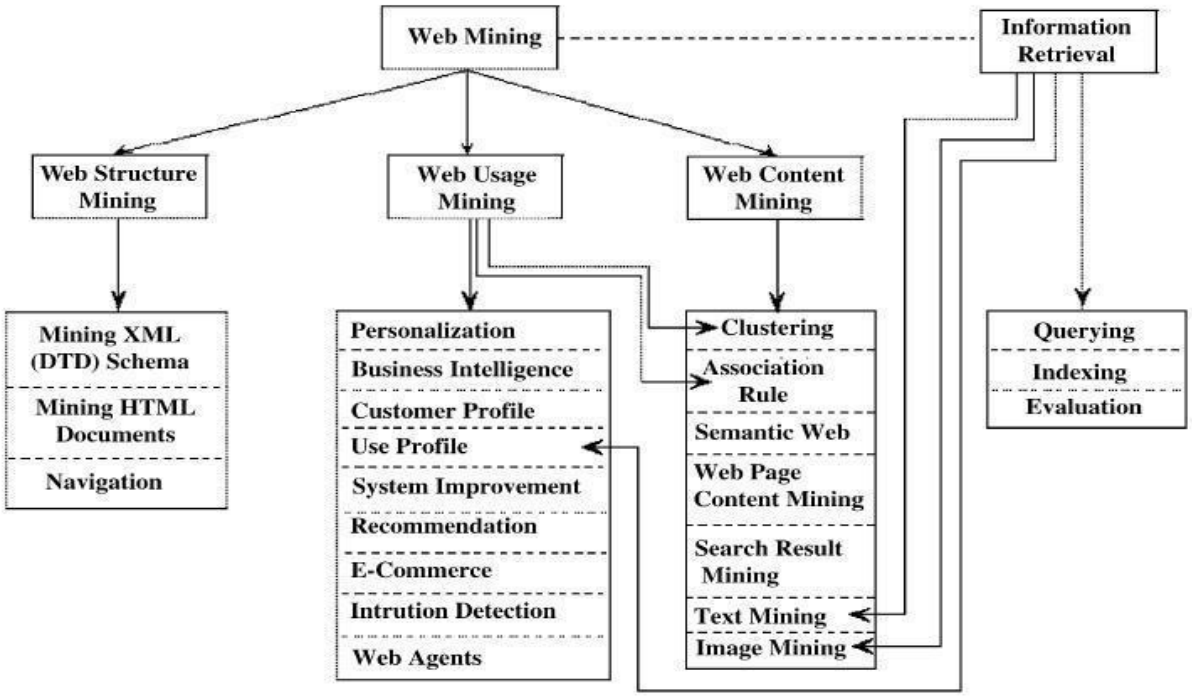

Figure 5: Web Mining Classification

Web Mining Analysis Techniques:

Following web mining analysis techniques can be added in the Web Mining section of the paper.

Association Analysis

Sequential Pattern Analysis

Clustering Analysis

Classification Analysis

Deviation Detection

Forecasting

Association Analysis

In a nutshell, Association analysis is finding interesting rules from large collections of data. This data mining technique is carried out using the association rules. Association rule mining has a wide range of applicability such Market basket analysis, Medical diagnosis/ research, Website navigation analysis, Homeland security and many more. Association rules are used to identify relationships among a set of items in database. These relationships are not based on inherent

properties of the data themselves (as with functional dependencies), however rather based on co-occurrence of the data items.

Association Rule:

Association rules are shaped like implication A=>B, where A is a subset of I and B is a subset of I and A∩ B =. I= {i1,..., im} is a transaction items set of database. The rule A=>B is established in transaction set I, has support S, and confidence C, namely:

Support (A=>B) = P (A∪B)

Confidence (A=>B) =P (B/A)

Sequential Pattern Analysis

Sequential pattern analysis and association analysis are similar, but the emphasis point is to analyze the around sequence relations among data. It can discover that such as “the customer purchases commodity X, then purchases commodity Y, then purchases commodity Z in some period of time, namely the appearance frequency of sequence X >Y >Z is high” and so on knowledge in the database. Sequence pattern analysis need to find the rules that has appeared in more than a certain minimum percentage rate (threshold) during a minimum time. This rule will make appropriate adjustment with the change of the form.

Clustering analysis

The input of clustering analysis is a group of unclassified records, and these records should be divided into several unknown classifications in advance. Clustering analysis divides the record set reasonably, determines each record in the classification through analysis of this data records in the database according to certain classification rule. The clusters generated by clustering are the set of a group of data objects. The similarity of these objects with other objects in the same clusters should be as far as possible big, and with other objects in the different clusters should be as far as possible small. Clustering can help the market analyst to discover the different customer group from the customer fundamental information warehouse,

and analyze the characteristic of each group. Clustering analysis can be used for data preprocessing of other algorithmic.

Classification analysis

Supposes a database and a set of classifications with different characteristics, first, classification analysis endows a tag for each record, then checks these marked records and describes them. Classification generally represents adopting rules or decision tree model, which can map the tuple in the database to one in a given classification. Class can be obtained with training data set (samples made of the tuples and the classifications in known database) by the related algorithm.

Deviation Detection

There are many abnormal conditions in database. It is necessary to find them from the data analysis in order to cause people's more concerns. The basic method of deviation detection is to look for the difference between the observation and the reference. The observation is often the value of certain domain or summary of more domain value. The reference is the forecasting of a given model, the external provided standard or another observation.

Forecasting

Forecasting utilizes the historical data to find out the change law, builds the model to forecast the kind, characteristic of the future data, etc. The typical method is regression analysis, which utilizes a large number of historical data, regard time as the variable to set up linear or non-linear regression equation. While forecasting, only need to input arbitrary time value, it can get the state of this time through regression equation.

Fertile Area of Research : Semantic Web & Web Mining

- With the revolutionary invention of World Wide Web (WWW) by Mr. Tim Berners Lee, we are experiencing the unprecedented and exponential growth in information sources available online specifically in the healthcare sector

- The Web Mining & Semantic Web research is at the cross road of research from several research communities such as database (OLTP & OLAP), Information Retrieval, Machine Learning, Natural Language Processing and Advance Textual Analytics.
- The research in this area would facilitate: Web structure optimization, Personalized customer service, describe web usage, realize Business Intelligence (BI), finding potential customers, intelligent web advertisements and Network Security as well.
- From the user's perspective extraction, filtration and evaluation of desired information and resources is becoming immensely important and beneficial.
- The success and growth of E-Healthcare E-Commerce would reach to the next level with the incorporation of Web Semantics & Web Mining techniques. Through this entire community of Web users in the world would be benefitted immensely. The rise and promotion of this version of E-Healthcare and E-commerce would bring world market under one unified umbrella of Web. This would also save time and efforts of millions of users.

Conclusion

We learned that most data on the web are so much unstructured that they can only be understood by humans; however the amount of data is so huge that they can only be processed efficiently by machines. This paper has given us an approach to solve such issues by the combination of Semantic Web and Web Mining techniques. We discussed how Semantic Web Mining can improve the results of Web Mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques.

Day by day the web and its usage are growing continuously, so does the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years has seen the emergence of web mining as a rapidly growing area, due to the efforts of the research

community as well as various organizations that are practicing. While concluding it would be worthy to describe the pros and cons of Web Mining.

Pros:

Web mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled ecommerce to do personalized marketing, which eventually results in higher trade volumes. The government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of the mining application can benefits the society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

Cons:

The most criticized ethical issue involving web mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user's interests. The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site. This trend has increased the amount of data being captured and traded increasing the likeliness of one's privacy being invaded. Some mining algorithms might use controversial

attributes like sex, race, religion, or sexual orientation to categorize individuals. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes. This process could result in denial of service or a privilege to an individual based on his race, religion or sexual orientation

References

[1] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In I. Horrocks and 1. Hendler, editors, Proceedings of the1st International Semantic Web Conference (ISWC-02), pages 264-278. Springer-Verlag, 2002.

[2] C.C. Aggarwal, Collaborative crawling: Mining user experiences for topical resource discovery, in: D. Hand, D. Keim, R. Ng (Eds.), Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2002, pp. 423–428

[3] Zaidi, Abidi, Manickam, "Distributed Data Mining From Heterogeneous Healthcare DataRepositories: Towards an Intelligent Agent-Based Framework", Proceedings of the 15 th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)

[4] Yuefeng Li, Web mining model and its applications for information gathering, Knowledge-Based Systems 17 (2004) 207–217, Elsevier

[5] Han, J., Pei , J. ,and Yin ,y. . "Mining Frequent Patterns without Candidate Generation : A Frequent- Pattern Tree Approach". In Proc. ACM-SIG MOD Int. Conf. Management of Data(SIG MOD'04) , pp. 53-87, 2004.

[6] Juan M. Santos, Luis Anido and Martín Llamas, Design of a Semantic Web-based Brokerage Architecture for the E-learning Domain. A Proposal for a Suitable Ontology, 35th ASEE/IEEE Frontiers in Education Conference, 2005 IEEE

[7] Gerd Stumme, Andreas Hotho, Bettina Berendt, Semantic Web Mining State of the art and future directions, Web Semantics: Science, Services and Agents on the World Wide Web 4 (2006) 124–143

[8] Han & Kamber, Data mining concepts and techniques, second edition, 2006, elsevier inc.

[9] Shmueli, Patel, Bruce, data mining for business intelligence, 2007 wiley inc.

[10] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang, Ontology-Based Automatic Classification and Ranking for Web Documents, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2007 IEEE

[11] Johannes K. Chiang, “Multidimensional Data Mining of Association Patterns in various Granularities for Healthcare Service Portfolio Management”, Proceedings of the 2007 IEEE IEEM

[12] Cheng Yeh, Applications of web mining for marketing of online bookstores, Expert Systems with Applications 36 (2009) 11249–11256, Elsevier Ltd.

[13] Dieter Fensel, Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce

[14] Li Zhu , Qing Yang , Wei Chen, Research on Ontology Integration Combined with Machine learning, Second International Conference on Intelligent Computation Technology and Automation, 2009 IEEE

[15] Bernard Suteja, Implementing Ontology for E-learning System Development in Indonesia based on Web 2.0, International Journal of Web Application, March 2009

[16] Pieter Noordhuis, Mining Twitter in the Cloud: A Case Study, 2010 IEEE 3rd International Conference on Cloud Computing

[17] Shen zihao, Research on E-commerce Application Based on Web Mining, 2010 International Conference on Intelligent Computing and Cognitive Informatics, 2010 IEEE

[18] Andreas S. Rath, An Ontology Based Approach for Detecting Knowledge Intensive Tasks, Journal of Digital Information Management, February 2011